Foundations > Maximum Likelihood Estimation of Gaussian Parameters

Maximum Likelihood Estimation of Gaussian Parameters

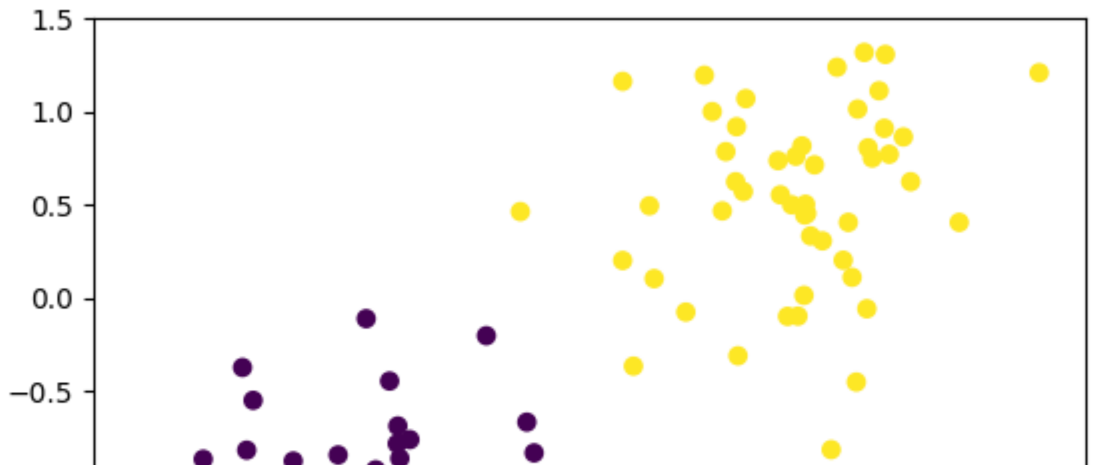
This text appearred in this post but because the math formatting disappeared from the original site, it is shamelessly copied here. Please note that in our notation N is m (number of examples / samples) and n is i (index of each example / sample).

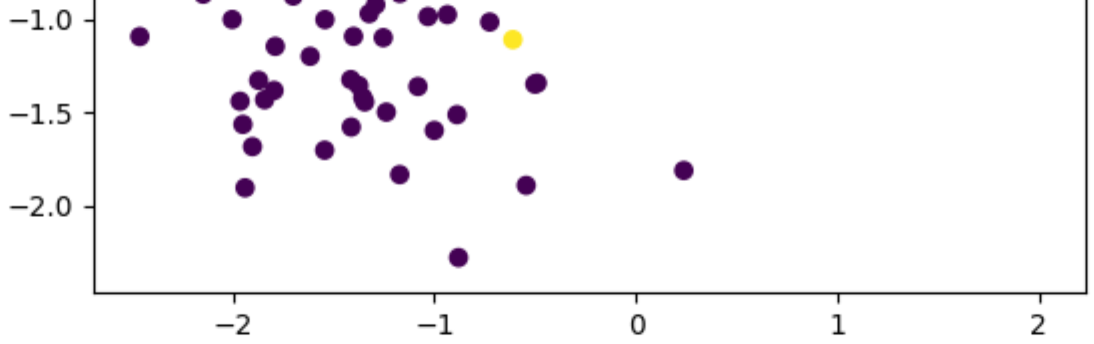
The Big Picture

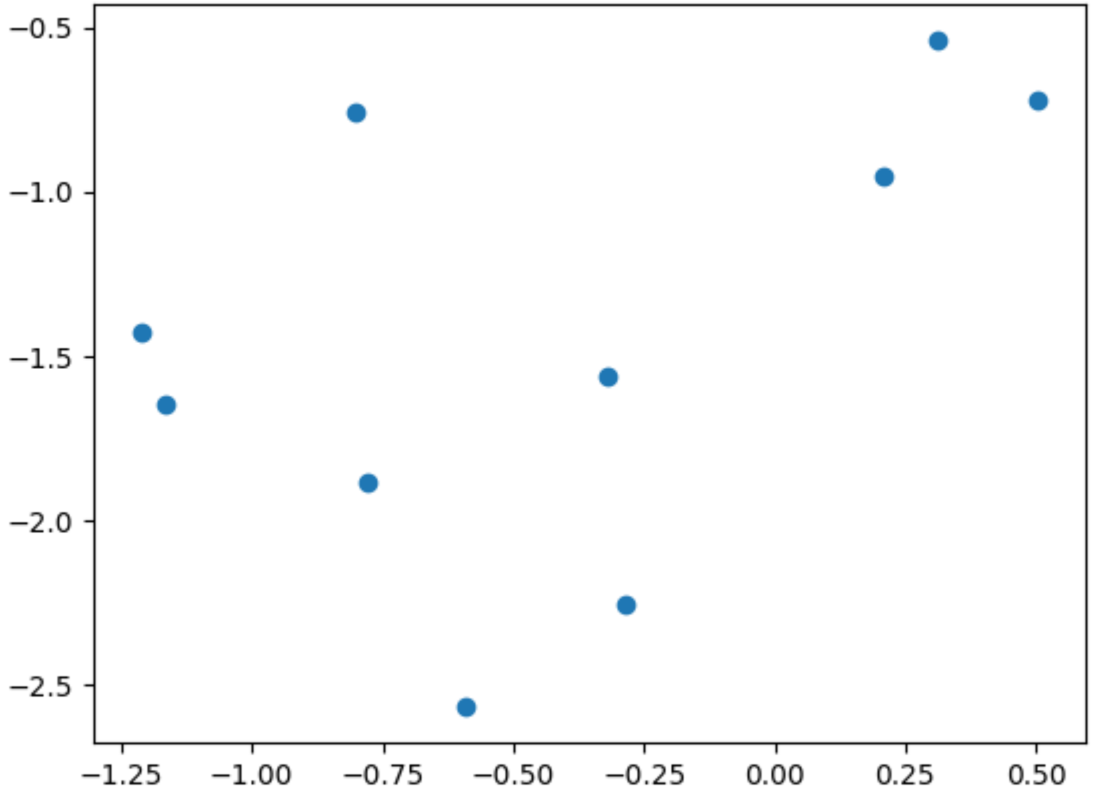
Maximum Likelihood Estimation (MLE) is a tool we use in machine learning to acheive a *very* common goal. The goal is to create a statistical model, which is able to perform some task on *yet unseen data*.

The task might be classification, regression, or something else, so the nature of the task does not define MLE. The defining characteristic of MLE is that it uses *only existing data* to estimate parameters of the model. This is in contrast to approaches which exploit *prior knowledge* in addition to existing data.[1]

Today, we're talking about MLE for Gaussians, so this is going to be a classification task. That is, we have data with labels, and we want to take some new data, and classify it *using the labels from the old data*. In the below images, we see data with labels (left), and new, unlabeled data (right). We want to be able to categorize each point from the `new data` as belonging to either the `purple` group or the `yellow` group.







Labeling dots as either `purple` or `yellow` sounds pretty boring, but the same idea applies to labeling emails as `spam` or `ham` or classifying audio clips as the vowel `[a]` or the vowel `[o]`.

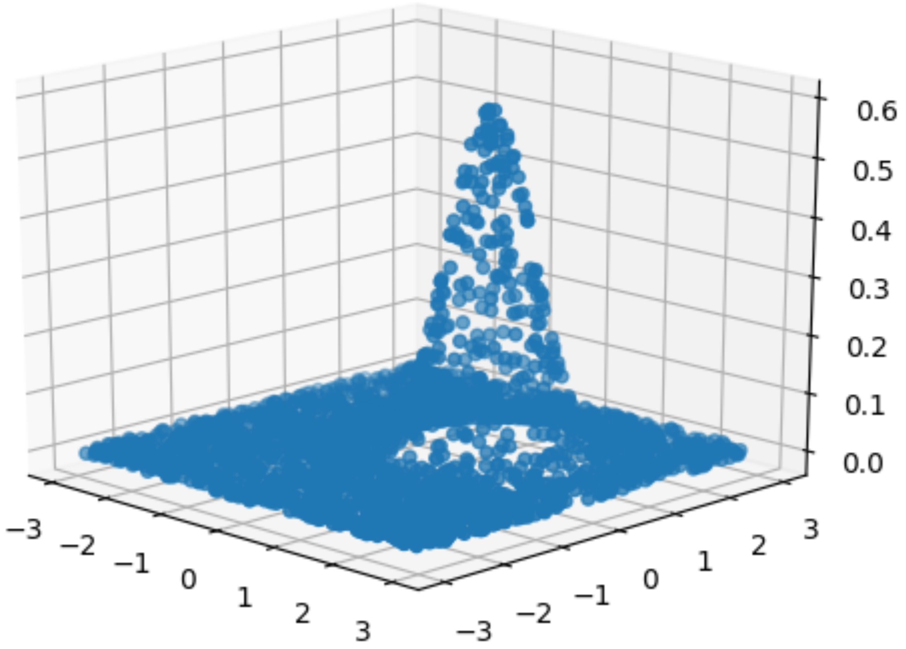
To make this post more tasty, let's pretend we're classifying `skittles` as `purple` or `yellow`.[2] We're classifying these skittles based on two dimensions `[x,y]`. Let's say the skittles have been rated by expert skittle-sommeliers on two traits: `x = aromatic lift` and `y = elegance`.

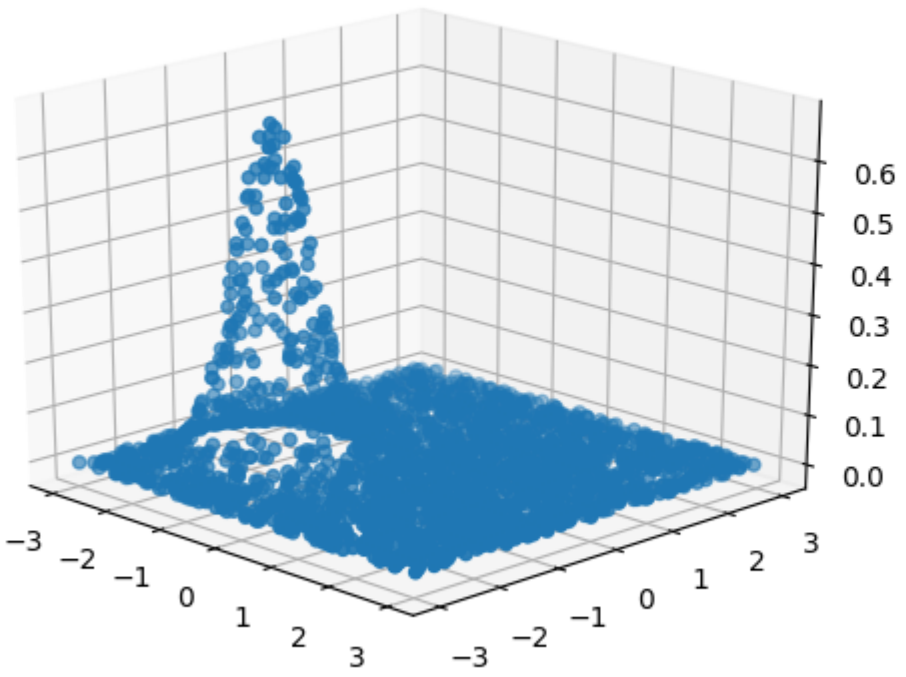
As you can see, `purple skittles` have bad ratings on both aromatic lift and elegance, whereas `yellow skittles` have been highly rated on both traits. Since these ratings are from expert skittle-sommeliers, they must be true.

To get our new, unlabeled data, we've given some new skittles to our expert sommeliers in a blind taste test. That is, the experts don't know what they ate, and neither do we. The only information available for each skittle is its rating on `aromatic lift` and `elegance`.

Now, we want to take ratings for each `mystery skittle` and figure out if it was a `purple skittle` or `yellow skittle`. To accomplish this task, we build a statistical model, learning its shape from the old ratings (i.e. the labeled data).

For the above data we can build two models (i.e. 2-D Gaussians), a `purple skittle model` and a `yellow skittle model`, and then see which model is more similar to a new rating on a `mystery skittle`. Another approach would be to build a single model (eg. a neural net) that distinguishes `purple skittles` from `yellow skittles`, and then see how it categorizes each `mystery skittle`.[3] Here, we're working with the former approach (build two models and see which one fits better).





We assume the data was in a sense *generated* by some process, and we're trying to model what that process was. This is called the generative approach. The model we're trying to learn is an approximation of the underlying process that created the data in the first place. So our data is just a sample from a process, and we want to learn the process.

At the end of the day, once we have our two models, we will use them to find which model was more likely to have *generated* the new data point. To take the leap from data $\rightarrow$ model, we need to not only estimate possible parameters of the model (eg. for Gaussians we need $[\mu, \Sigma]$), but we want the *best* model possible for our data. That's where MLE comes into play.

MLE as Parameter Estimation

MLE is one flavor of parameter estimation in machine learning, and in order to perform parameter estimation, we need:

1. some data $\mathbf{X}$
2. some hypothesized generating function of the data $f(\mathbf{X}, \theta)$
3. a set of parameters from that function θ
4. some evaluation of the goodness of our parameters (an objective function)

In MLE, the objective function (evaluation) we chose is the *likelihood* of the data given our model. This intuitively makes sense if you keep in mind that we don't get to change our data, and we have to make some assumption about the form of our model, but we *can* adjust the parameterization of our model. So, we are limited to adjusting θ, and we might as well choose the best θ for our data. To find the best θ then, we need to find the θ which maximizes our evaluation function (the likelihood). Therefore, in its general form the MLE is:

$$\theta_{MLE} = \operatorname*{argmax}_{\theta} p(\mathbf{X}|\theta)$$

Likelihood for a Gaussian

We assume the data we're working with was generated by an underlying Gaussian process in the real world. As such, the likelihood function ($\mathcal{L}$) is the Gaussian itself.

$$\begin{aligned} \mathcal{L} = p(\mathbf{X}|\theta) &= \mathcal{N}(\mathbf{X}|\theta) \\ &= \mathcal{N}(\mathbf{X}|\mu, \Sigma) \end{aligned}$$

Therefore, for MLE of a Gaussian model, we will need to find good estimates of both parameters: μ and Σ:

$$\begin{aligned} \mu_{MLE} &= \operatorname*{argmax}_{\mu} \mathcal{N}(\mathbf{X}|\mu, \Sigma) \\ \Sigma_{MLE} &= \operatorname*{argmax}_{\Sigma} \mathcal{N}(\mathbf{X}|\mu, \Sigma) \end{aligned}$$

Solving these two above equations to find the best μ and Σ is a job for our good old friends from calculus… partial derivatives!

Before we can get to the point where we can find our best μ and Σ, we need to do some algebra, and to make that algebra easier, instead of just using the likelihood function as our evaluation function, we're going to use the log likelihood. This makes the math easier and it doesn't run any risks of giving us worse results. That's because the $\log()$ function is monotonically increasing, and therefore us worse results. That's because the $\log()$ function is monotonically increasing, and therefore

$$\underset{\theta}{\operatorname{argmax}} \log(f(\theta)) == \underset{\theta}{\operatorname{argmax}} f(\theta)$$

So now, we know that we want to get the best parameters $\theta = [\mu, \Sigma]$ for a dataset $\mathbf{X}$ evaluating on a normal, Gaussian distribution.

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \log(\mathcal{N}(\mathbf{X}|\theta))$$

Since in reality our dataset $\mathbf{X}$ is a set of labeled data $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \ldots \mathbf{x}_n]$, to evaluate our parameters on the entire dataset, we need to sum up the log likelihood for each data point.

$$\log(\mathcal{N}(\mathbf{X}|\theta)) = \sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\theta))$$

Remember how that θ is a general catch-all for any set of parameters? Let's be more explicit with our Gaussian parameters $[\mu, \Sigma]$:

$$\sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\theta)) = \sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\mu, \Sigma))$$

Here we're going to make a big simplfying assumption (and in reality a pretty common one). We're going to assume that our Gaussians have diagonal covariance matrices. So the full covariance matrix $\mathbf{\Sigma}$ gets replaced by a diagonal variance vector σ^2:

$$\sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\mu, \Sigma)) = \sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\mu, \sigma^2))$$

Now, with this simplification, we can take a look at our fully specified log likelihood function that we'll be working with from here on out.

$$\sum_{n=1}^{N} \log(\mathcal{N}(\mathbf{x}_n|\mu, \sigma^2)) = \sum_{n=1}^{N} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp^{-\frac{1}{2}\left(\frac{(x_n-\mu)^2}{\sigma^2}\right)}\right)$$

Now we have the likelihood as we want it (Gaussian, logged, diagonal covariance matrix). Let's not forget what our main goal is! We want to find the best parameters for our model given our data, so we're going to find the $\underset{\mu}{\operatorname{argmax}}$ and $\underset{\sigma^2}{\operatorname{argmax}}$. Before we can get to that point, we need to do some simplifications to the log likelihood to make it easier to work with (that is, since we will soon be doing some partial derivatives, the log likelihood in its current form it will lead to some messy math). In the following, $\mathcal{LL}$ means *log likelihood*.

The next first steps take advantage of our choice to use the log likelihood instead of the plain likelihood. Our first step will be to use the log product rule:

$$\begin{aligned}\mathcal{LL} &= \sum_{n=1}^{N} \log\Big(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp^{-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2})}\Big) \\ &= \sum_{n=1}^{N}\Big(\log\big(\frac{1}{\sqrt{2\pi\sigma^2}}\big) + \log\Big(\exp^{-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2})}\Big)\Big)\end{aligned}$$

Now we will use the log quotient rule:

$$\begin{aligned}\mathcal{LL} &= \sum_{n=1}^{N}\Big(\log\big(\frac{1}{\sqrt{2\pi\sigma^2}}\big) + \log\Big(\exp^{-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2})}\Big)\Big) \\ &= \sum_{n=1}^{N}\Big(\log(1) - \log\big(\sqrt{2\pi\sigma^2}\big) + \log\Big(\exp^{-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2})}\Big)\Big)\end{aligned}$$

Now, we'll use the log power rule:

$$\begin{aligned}\mathcal{LL} &= \sum_{n=1}^{N}\Big(\log(1) - \log\big(\sqrt{2\pi\sigma^2}\big) + \log\Big(\exp^{-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2})}\Big)\Big) \\ &= \sum_{n=1}^{N}\Big(\log(1) - \log\big(\sqrt{2\pi\sigma^2}\big) + \Big(-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2}) \cdot \log(e)\Big)\Big)\end{aligned}$$

We're now going to be explicit that the $\log()$ function we used was base e. This allows us to simplify $\log_e(e) = 1$ as well as $\log(1) = 0$ (regardless of base).

$$\begin{aligned}\mathcal{LL} &= \sum_{n=1}^{N}\Big(\log(1) - \log\big(\sqrt{2\pi\sigma^2}\big) + \Big(-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2}) \cdot \log(e)\Big)\Big) \\ &= \sum_{n=1}^{N}\Big(-\log\big(\sqrt{2\pi\sigma^2}\big) + \Big(-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2})\Big)\Big)\end{aligned}$$

We can apply the power rule one more time (remember that $\sqrt{x} = x^{1/2}$).

$$\mathcal{LL} = \sum_{n=1}^{N}\Big(-\log\big(\sqrt{2\pi\sigma^2}\big) + \Big(-\frac{1}{2}(\frac{(x_n-\mu)^2}{\sigma^2})\Big)\Big)$$

$$= \sum_{n=1}^{N} \Big(-\frac{1}{2} \cdot \log(2\pi\sigma^2) - \frac{1}{2}\Big(\frac{(x_n-\mu)^2}{\sigma^2}\Big)\Big)$$

Now for some basic algebra simplification:

$$\begin{aligned}
\mathcal{LL} &= \sum_{n=1}^{N} \Big(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\Big(\frac{(x_n-\mu)^2}{\sigma^2}\Big)\Big) \\
&= -\frac{N}{2}\log(2\pi\sigma^2) + \sum_{n=1}^{N} -\frac{1}{2}\Big(\frac{(x_n-\mu)^2}{\sigma^2}\Big) \\
&= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2
\end{aligned}$$

Now we have our log likelihood function ($\mathcal{LL}$) in a nice, easy to work with form. Now we need to take the ($\mathcal{LL}$) and *estimate* it's *maximum* for our parameters. We've got the $\mathcal{LL}$ ready for $\underset{\theta}{\operatorname{argmax}}(\mathcal{LL})$, now we need to do the $\underset{\theta}{\operatorname{argmax}}$ part. This is where we get a little help from our friends, partial derivatives. We need *partial* derivatives because our θ is really two variables $[\mu, \sigma^2]$, and we need the best value for each.

So, now we're going to solve the problem for each variable one-by-one:

$$\underset{\mu}{\operatorname{argmax}}\, \mathcal{LL}(X|\mu,\sigma^2)$$
$$\underset{\sigma^2}{\operatorname{argmax}}\, \mathcal{LL}(X|\mu,\sigma^2)$$

To get the argmax for each parameter we have to do two things. First, we must:

1. derive the partial derivative of the function with respect to that parameter, and then
2. set that partial derivative to zero, and solve for our parameter

By doing step (1), we get an equation for the change of the function, and we know that if the function isn't changing at a certain point, that point is a maximum or a minimum (or a saddle point).

In our case, our log-likelihood function is concave, so we know that at $\frac{\partial \mathcal{LL}}{\partial \mu} = 0$ we get a maximum.

MLE of μ

First we'll work to solve for the mean of our Gaussian, μ. Remember we've got our likelihood function in a simple form:

$$\mathcal{LL} = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2$$

and now we want to get the best μ for that function:

$$\operatorname*{argmax}_{\mu} \mathcal{LL}(X|\mu, \sigma^2) := \frac{\partial \mathcal{LL}}{\partial \mu} = 0$$

So, to get to the point where we can set the partial derivative to zero and solve, we need to first find the partial derivative with respect to μ:

$$\frac{\partial \mathcal{LL}}{\partial \mu} = \frac{\partial}{\partial \mu}\Big(-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big)$$

Now let's start simplifying! First we can right off the bat get rid of the first term since it doesn't contain μ, and therefore is practically speaking a constant:

$$\begin{aligned}\frac{\partial \mathcal{LL}}{\partial \mu} &= \frac{\partial}{\partial \mu}\Big(-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big)\\ &= \frac{\partial}{\partial \mu}\Big(-\frac{N}{2}\log(2\pi\sigma^2)\Big) + \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big)\\ &= 0 + \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big)\\ &= \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big)\end{aligned}$$

Next, remember that the summation expression is just a convenient way to write a longer expression:

$$\sum_{n=1}^{N} f(x_n) = f(x_1) + f(x_2) + \ldots + f(x_N)$$

Also, We know from the summation rule that :

$$\frac{\partial}{\partial x}\big(f(x)+g(x)\big) = \frac{\partial}{\partial x}f(x) + \frac{\partial}{\partial x}g(x)$$

Therefore, when we take the derivative of a sum, we can reformulate it as a sum of derivatives:

$$\frac{\partial}{\partial x}\sum_{n=1}^{N} f(x_n) = \sum_{n=1}^{N}\frac{\partial}{\partial x}f(x_n)$$

Now, getting back to the problem at hand, we can move the derivative operator inside the summation term:

$$\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \mu} &= \frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)\\
&= \frac{\partial}{\partial \mu}\Big(\sum_{n=1}^{N}-\frac{1}{2\sigma^2}(x_n-\mu)^2\Big)\\
&= \sum_{n=1}^{N}\frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}(x_n-\mu)^2\Big)
\end{aligned}$$

Now we can use the product rule:

$$\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \mu} &= \sum_{n=1}^{N}\frac{\partial}{\partial \mu}\Big(-\frac{1}{2\sigma^2}\cdot(x_n-\mu)^2\Big)\\
&= \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \mu}\big(-\frac{1}{2\sigma^2}\big)\cdot(x_n-\mu)^2 + \big(-\frac{1}{2\sigma^2}\big)\cdot\frac{\partial}{\partial \mu}(x_n-\mu)^2\Big)
\end{aligned}$$

Now some terms will nicely drop out:

$$\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \mu} &= \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \mu}\big(-\frac{1}{2\sigma^2}\big)\cdot(x_n-\mu)^2 + \big(-\frac{1}{2\sigma^2}\big)\cdot\frac{\partial}{\partial \mu}(x_n-\mu)^2\Big)\\
&= \sum_{n=1}^{N}\Big(0 + \big(-\frac{1}{2\sigma^2}\big)\cdot\frac{\partial}{\partial \mu}(x_n-\mu)^2\Big)\\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\frac{\partial}{\partial \mu}(x_n-\mu)^2
\end{aligned}$$

At this point we can use the chain rule, $\frac{\partial}{\partial x}\big(f(g(x))\big) = \frac{\partial}{\partial x}f(g(x))\cdot\frac{\partial}{\partial x}g(x)$, with $g(x)=(x-\mu)$ and $f(x)=x^2$.

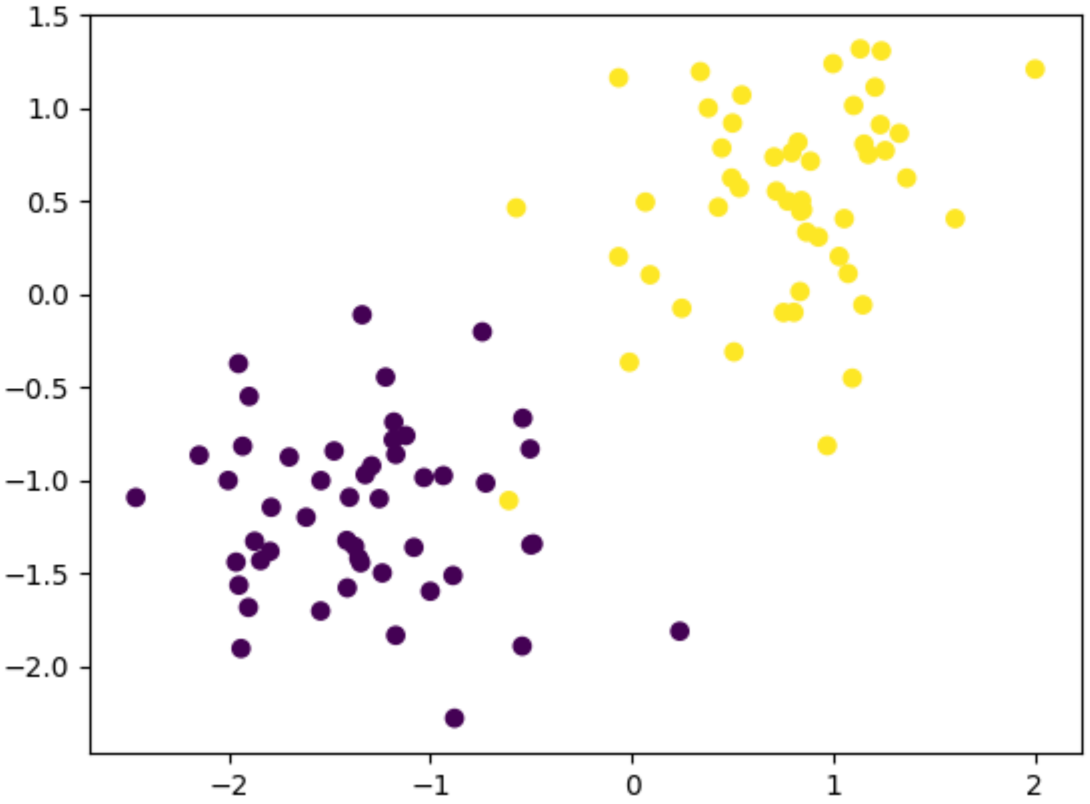
$$\begin{aligned}\frac{\partial \mathcal{LL}}{\partial \mu} &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}\frac{\partial}{\partial \mu}(x_n-\mu)^2 \\ &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}2(x_n-\mu)\cdot -1 \\ &= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)\end{aligned}$$

Yay! We've done as much simplifying as we can at this point, and gotten rid of all of our $\frac{\partial \mathcal{LL}}{\partial \mu}$ terms!

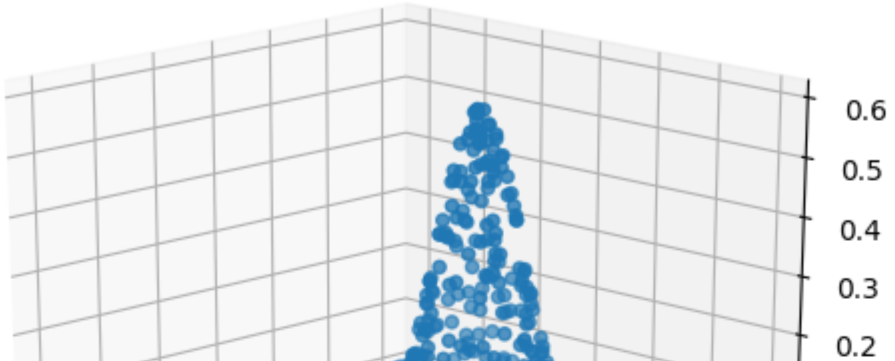
Now what we have is the simplest form of the partial derivative of our likelihood function with respect to μ. Now we want to use this equation to find the best μ, so we set it equal to zero, and solve for μ.

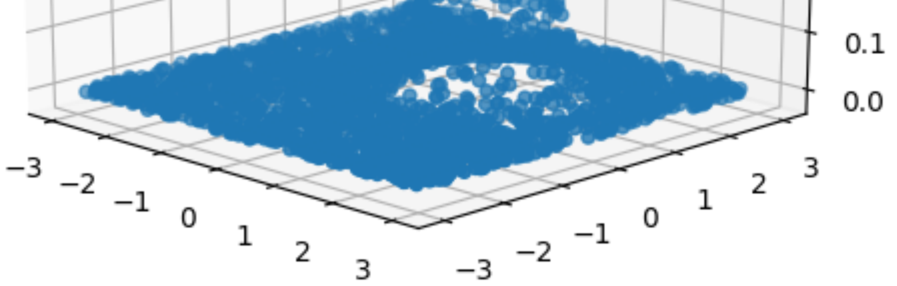
$$\begin{aligned}\frac{\partial \mathcal{LL}}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) \\ 0 &= \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu) \\ 0 &= \sum_{n=1}^{N}(x_n-\mu) \\ 0 &= \sum_{n=1}^{N}x_n - \sum_{n=1}^{N}\mu \\ 0 &= \sum_{n=1}^{N}x_n - N\cdot\mu \\ N\cdot\mu &= \sum_{n=1}^{N}x_n \\ \mu &= \frac{1}{N}\sum_{n=1}^{N}x_n\end{aligned}$$

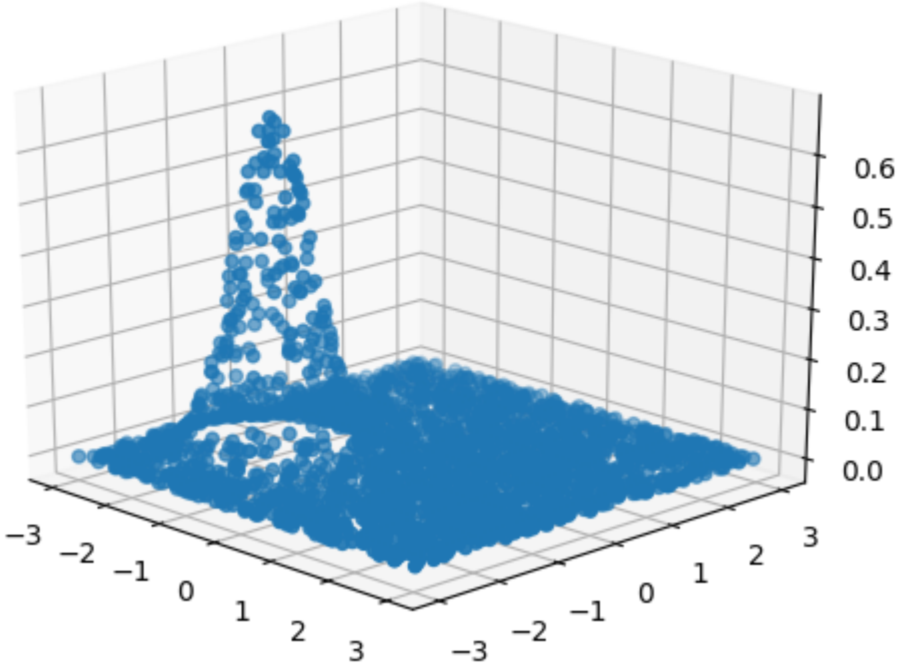
Huzzah! We've reached the promised land! We now have a formula we can use to estimate one model parameter (μ) from our data ($\mathbf{X}$). Let's take a second to think about what this formula means. Remember that we started with a bunch of data points:



We want to take that data and make some models which we think represent that data well. In our case we are learning two Gaussian models, one for the `yellow skittle` data and one for the `purple skittle` data:







To make our Gaussians fit the data as well as we can, we can do two things: (1) move the center of the curve or (2) adjust the width of the peak. Right now, with μ we're only talking about the placement of the center of the curve, we're not talking at all about its width.

Where is the best place to put a bell curve to cover all our data? Well, how about the center of our data! Dead-center, bull's eye, whatever you call it, we're putting our Gaussian right in the middle of it all (the

Dead center, bulls eye, whatever you call it, we're putting our Gaussian right in the middle of it all (the mean). We're taking all our points, summing them up, and dividing by the number of data points. This is the *average*, and it is (for the likelihood function) the best place to put the mean of our model.

Getting back to the skittles, the center of the curve $[\mu_x, \mu_y]$ for our `yellow skittle model` comes directly from our sommeliers' ratings (i.e. $[\mu_x]$ = average rating for `aromatic lift` and $[\mu_y]$ = average rating for `elegance`). Our `purple skittle model` was centered in the exact same way.

Sure enough, if you take a look at the data, you'll see that the `yellow skittle` data is grouped around the point `[-1,-1]` and that the `purple skittle` data points are all clustered around `[1,1]`. Now take a look at the models we've made. You'll see that the center of the peak for the `purple skittle model` is somewhere near `[-1,-1]` and that the peak of the `yellow skittle model` is around `[1,1]`.

MLE of σ^2

Now let's tackle the second parameter of our Gaussian model, the variance σ^2!

$$\begin{aligned}\frac{\partial \mathcal{LL}}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\Big(-\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\ &= \frac{\partial}{\partial \sigma^2}\Big(-\frac{N}{2}\log(2\pi\sigma^2)\Big) + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)\end{aligned}$$

Let's start with the product rule for the lefthand term:

$$\begin{aligned}\frac{\partial \mathcal{LL}}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\Big(-\frac{N}{2}\log(2\pi\sigma^2)\Big) + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\ &= \frac{\partial}{\partial \sigma^2}\big(-\frac{N}{2}\big)\cdot\log(2\pi\sigma^2) + \big(-\frac{N}{2}\big)\cdot\frac{\partial}{\partial \sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\ &= 0 + \big(-\frac{N}{2}\big)\cdot\frac{\partial}{\partial \sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\ &= -\frac{N}{2}\cdot\frac{\partial}{\partial \sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)\end{aligned}$$

Now we can use the chain rule for our term with the log operator, $\frac{\partial}{\partial x}\big(f(g(x))\big) = \frac{\partial}{\partial x}f(g(x)) \cdot \frac{\partial}{\partial x}g(x)$, with $g(x) = 2\pi x$ and $f(x) = log(x)$.

$$\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \sigma^2} &= -\frac{N}{2} \cdot \frac{\partial}{\partial \sigma^2}\big(\log(2\pi\sigma^2)\big) + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\
&= -\frac{N}{2} \cdot \frac{1}{2\pi\sigma^2} \cdot 2\pi + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\
&= -\frac{N}{2} \cdot \frac{1}{\sigma^2} + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\
&= -\frac{N}{2\sigma^2} + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)
\end{aligned}$$

Now, using the same logic as above with μ, we can move the derivative operator inside the summation operator:

$$\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \frac{\partial}{\partial \sigma^2}\Big(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(-\frac{1}{2\sigma^2}(x_n-\mu)^2\big)\Big)
\end{aligned}$$

And again, the product rule:

$$\begin{aligned}
\frac{\partial \mathcal{LL}}{\partial \sigma^2} &= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(-\frac{1}{2\sigma^2}(x_n-\mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(-\frac{1}{2\sigma^2}\big) \cdot (x_n-\mu)^2\big) + \big(-\frac{1}{2\sigma^2}\big) \cdot \frac{\partial}{\partial \sigma^2}(x_n-\mu)^2\big)\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(-\frac{1}{2\sigma^2}\big) \cdot (x_n-\mu)^2\big) + 0\Big) \\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(-\frac{1}{2\sigma^2}\big) \cdot (x_n-\mu)^2\big)\Big)
\end{aligned}$$

Now let's be careful with our exponents, since we're taking the derivative of the function with respect to a squared variable σ^2:

$$\frac{\partial \mathcal{LL}}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial \sigma^2}\big(-\frac{1}{2\sigma^2}\big) \cdot (x_n-\mu)^2\big)\Big)$$

$$= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial\sigma^2}\big(-\frac{1}{2}\cdot\sigma^{-2}\big)\cdot(x_n-\mu)^2\big)\Big)$$

Now it's obvious that we need the product rule:

$$\begin{aligned}
\frac{\partial\mathcal{LL}}{\partial\sigma^2} &= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial\sigma^2}\big(-\frac{1}{2}\cdot\sigma^{-2}\big)\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{\partial}{\partial\sigma^2}\big(-\frac{1}{2}\big)\cdot\sigma^{-2} + \big(-\frac{1}{2}\cdot\frac{\partial}{\partial\sigma^2}\sigma^{-2}\big)\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(0 + \big(-\frac{1}{2}\cdot\frac{\partial}{\partial\sigma^2}\sigma^{-2}\big)\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(-\frac{1}{2}\cdot\frac{\partial}{\partial\sigma^2}\sigma^{-2}\cdot(x_n-\mu)^2\big)\Big)
\end{aligned}$$

Now, we're going to use the chain rule again, by first treating $\sigma^{-2} = (\sigma^2)^{-1}$, then we can see $f(g(x))$ with $f(x) = x^{-1}$ and then $g(x) = x^2$:

$$\begin{aligned}
\frac{\partial\mathcal{LL}}{\partial\sigma^2} &= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(-\frac{1}{2}\cdot\frac{\partial}{\partial\sigma^2}\sigma^{-2}\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(-\frac{1}{2}\cdot\frac{\partial}{\partial\sigma^2}\big((\sigma^2)^{-1}\big)\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(-\frac{1}{2}\cdot -1\cdot(\sigma^2)^{-2}\cdot 1\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{1}{2}\cdot(\sigma^2)^{-2}\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\Big(\frac{1}{2\sigma^4}\cdot(x_n-\mu)^2\big)\Big)\\
&= -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2\\
&= \frac{1}{2\sigma^2}\Big(-N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\Big)
\end{aligned}$$

Huzzah! We've gotten our partial derivative for $\mathcal{LL}$ with respect to σ^2as simplified as we can. Now let's find the best $\sigma^2$for our data by setting the equation equal to zero and solving for σ^2

find the best σ^2 for our data by setting the equation equal to zero and solving for σ^2:

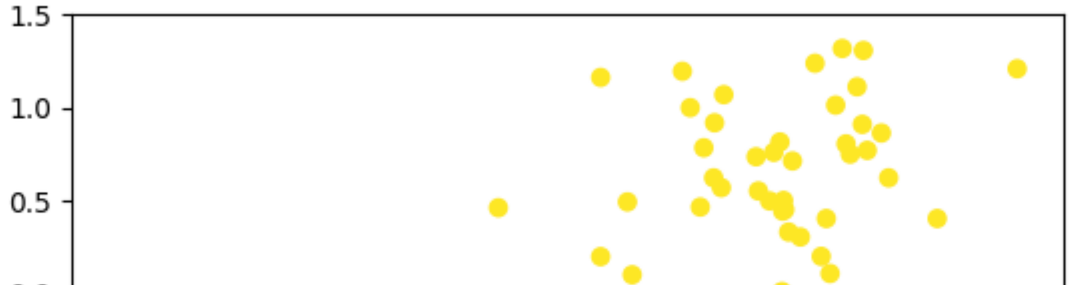
$$
\begin{aligned}
0 &= \frac{1}{2\sigma^2}\Big(- N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\Big) \\
0 &= -N + \frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 \\
N\sigma^2 &= \sum_{n=1}^{N}(x_n - \mu)^2 \\
\sigma^2 &= \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2
\end{aligned}
$$

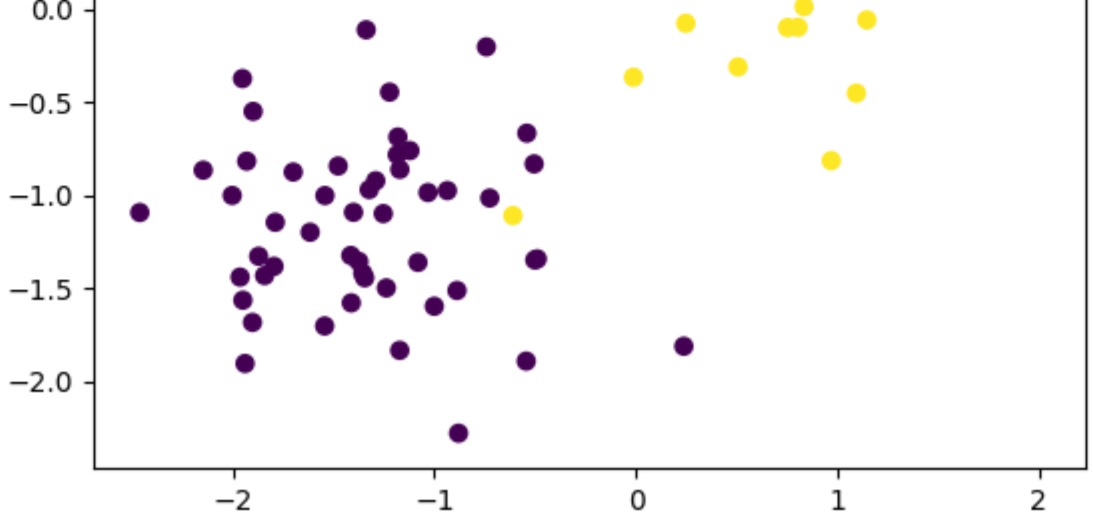
And there we have it! All that work has boiled down to a simple equation do get the best σ^2 for our Gaussian given our data. Just like with μ if we take a second to look at the equation, we find it has a very intuitive interpretation.

We're iterating over each data point (x_n) and finding its particular deviation from the mean of all the data points μ. We sum up all the deviations, and then take the average!

Just as when we were finding the best μ for our Gaussian by setting it to the average (μ) of the data, we're now setting our standard deviation to be, well, the *standard* deviation of the data! Think of *standard* as being a synonym to *average*, and it becomes pretty clear.

Thinking back to the `skittles`, what we've done here is taken each skittle, one-by-one, and figured out how far it deviates from the mean on a certain rating. For example, we know that on average, our expert sommeliers rated `purple` skittles to have a `[-1]` score for `elegance`. However, not every `purple skittle` got a score of `[-1]`. Each skittle deviated from that average, and if we add up how much each skittle deviated (after squaring), we with the average deviation. Take a look again at our skittle data:





Specifically, look at the `purple skittles`. You see the purple skittle that got a rating of about `[0,-2]`? That skittle had pretty horrible `elegance` and better-than-average `aromatic lift`. This skittle obviously *deviated* from the normal rating. We take this skittle along with every other, calculate that deviation, and get our $\sigma^2$for our model.

Conclusion

We did a lot of algebra, some calculus, and used some tricks with $\log_e$ to get to this point.

Along the way it's easy to get lost in the weeds, but if we keep in mind that all these equations have some interpretation, we can catch the big picture. In our case, the big picture is very clear:

> When using Maximum Likelihood Estimation to estimate parameters of a Gaussian, set the mean of the Gaussian to be the mean of the data, and set the standard deviation of the Gaussian to be the standard deviation of the data.

$$\mu_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{n=1}^{N} (x_n - \mu)^2$$

I hope this was helpful or interesting! If you find errors or have comments, let me know!

Footnotes

1. For another approach to parameter estimation using not only information from the data, but a prior bias, see Maximum A Posteriori estimation. ↩
2. I have no sponsorship from Skittles, the Wrigley Company, or Mars, Inc. All the views expressed in this post are my own. However, if they would like to throw some cash my way, I would not be upset. ↩
3. This is an example of a discriminative model, as opposed to a generative model. The classification approach described here is the latter approach. ↩